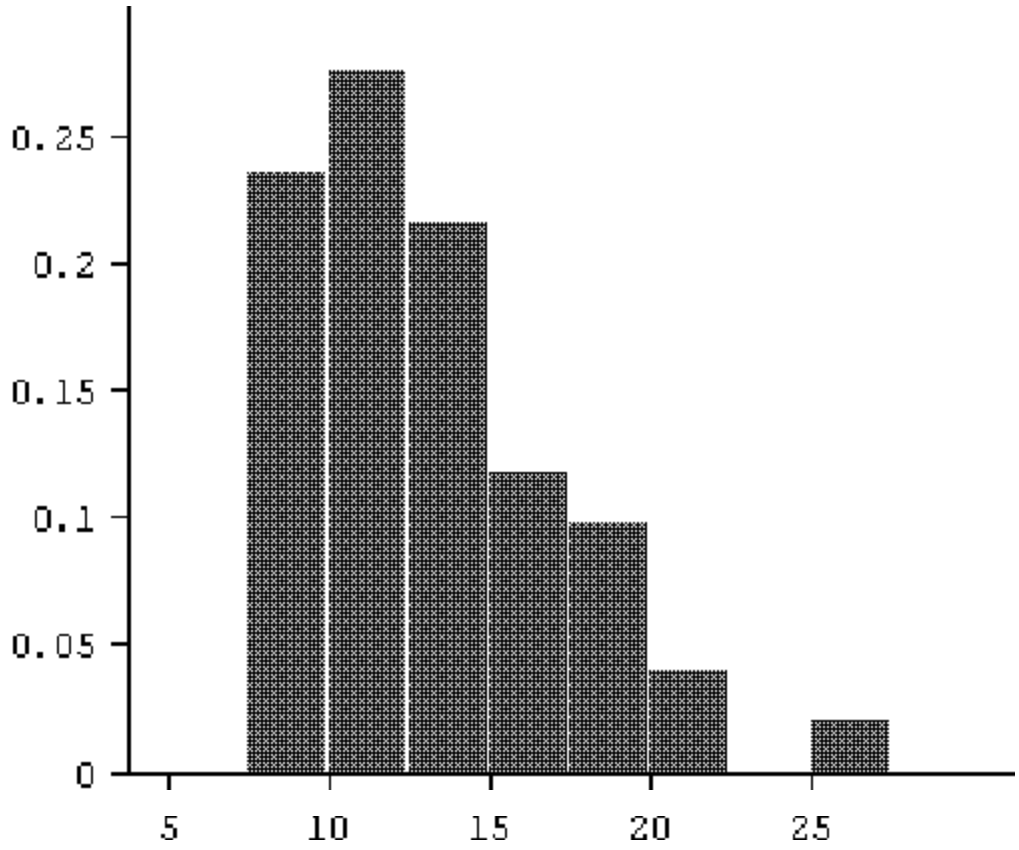
Ch. 3: Histograms

If we have a large amount of data that we wish to understand, a good start is to use the information to draw a simple plot or graph. This idea is often called exploratory data analysis.

If we are just interested in the distribution of a single set of numbers, we could use a _____.


0.25
0.2
0.15
0.1
0.05
0
5
10
15
20
25

Ex: The following data set shows the age of people watching a recent Disney movie in a local cinema (there was no age restriction for children).

Viewers Age in Years						
13	12	11	19	24	2	13
17	15	2	17	15	7	15
13	27	4	16	13	9	5
8	19	4	17	12	5	28
7	23	13	13	6	21	20
10	6	10	7	17	18	19
10	2	13	9	27	17	14
21	9	19	12	3	18	11
18	11	25	11	10	12	14
17	5	14	30	7	15	4
19	18	11	19	1	13	8
15	20	4	4	14	13	10
15	24	14	11	22	15	7
23	15	12	18	16	6	23
12	14	23	18	10	25	18
24						

What can we conclude from these numbers (without any further work)?

We can do better by constructing a _____.

We break our set of possible values into _____ and count how many data values fall into each.

Viewers Age Distribution Table

Years	Count	Percent
0-4	10	9
5-9	16	15
10-14	33	31
15-19	29	27
20-24	12	11
25-29	5	5
30-34	1	1
	106	99

This is better, but it's still hard to grasp the entire situation at once.

We can do better still by displaying the data graphically in a *histogram*.

The area of each block is proportional to the number of data points in that class interval.

- When just looking at the histogram (and not at the distribution table), we would observe that about 25% of the viewers ages were from 15 to 19 years.

- Which percentage of viewers is in the 20 to 24 years age group?

- And which percentage in the 0 to 9 years age group?

In this example, the heights in the histogram are directly proportional to percentage, because the class intervals all have equal width. This might sometimes not be the case!

Constructing a Histogram

1. Start with a *distribution table*.

2. Set up the *horizontal axis*, making sure spacing is consistent.

3. *Calculate heights*. Since

   $$\text{percentage} = \text{area} = \text{width} \times \text{height},$$

   we must calculate the height of each block as

   $$\text{height} = \text{percentage}/\text{width}.$$

4. *Draw the blocks* using the heights just calculated.

Ex: The 1990 Census found the following breakdown of years of education for adults over 25 in Cache County:

Years of Education				
at least	but less than	Percentage	Width	Height
0	9	3		
9	12	8		
12	13	25		
13	16	34		
16	17	19		
17	23	11		

Histograms that use percentages on the vertical axis are drawn using the _____.

In the density scale, area represents percentage, height represents crowding, i.e., the percentage per horizontal unit.

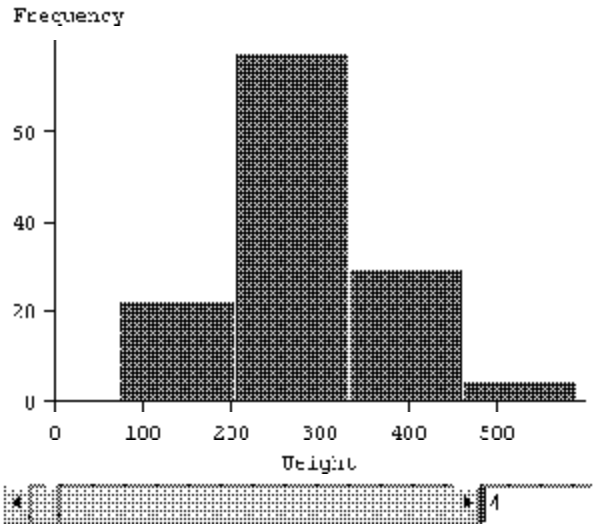
About what percentage of the over-25 population of Cache County in 1990 had at least 12 but less than 15 years of education?

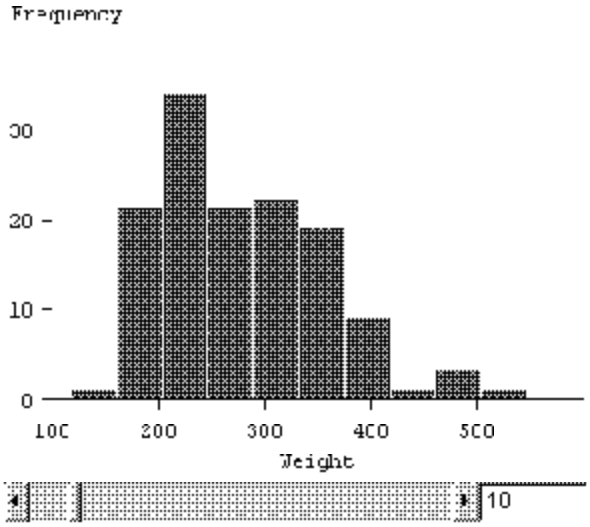
With the density scale, areas of blocks are percentages. The area under the histogram over an interval equals the percentage of cases in that interval. The total area is 100%.

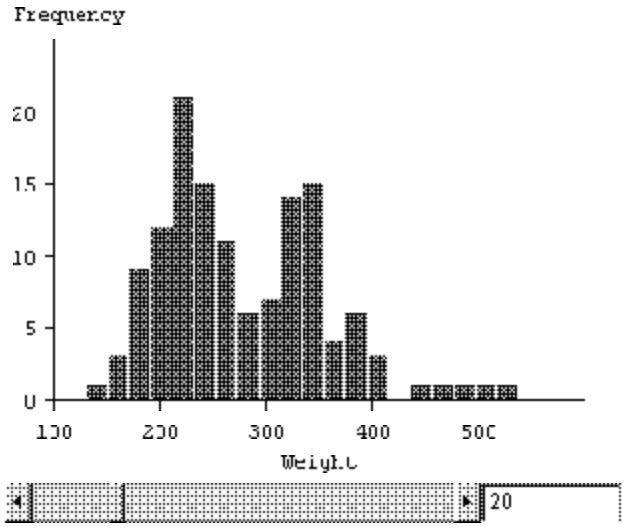
Warning:

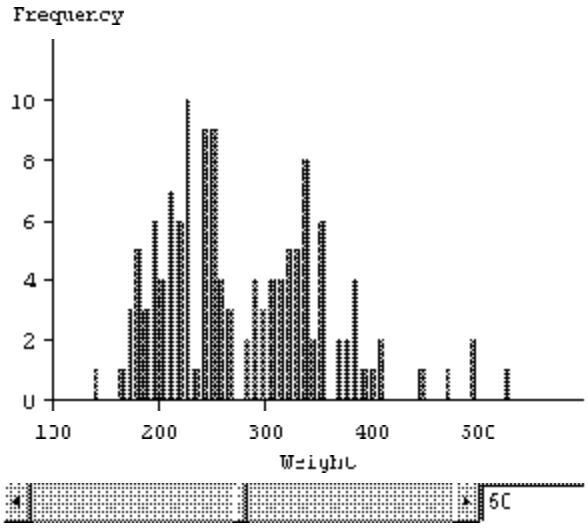
The visual impression we get from a histogram is highly depending on the width of each class and the starting point of each class. Even worse, histograms with different class widths could easily provide a misleading visual impression!

Following are 4 histograms of the same data set, the weights in pounds of 132 professional male athletes. What can you conclude about the distribution of the weights when looking at each of the 4 histograms?


Frequency
50
40
20
0
0
100
200
300
400
500
Weight
1


Frequency
30
20
10
0
100
200
300
400
500
Weight
10


Frequency
20
15
10
5
0
100
200
300
400
500
Weight
20


Frequency
10
8
6
4
2
0
100
200
300
400
500
Weight
50

Now imagine what could happen if we also modify the width of individual classes, allowing classes with different widths.

Recommendation: When constructing histograms (and distribution tables), use *equal class widths* whenever possible!